Proceedings Paper:

Learning Functional Object-Categories from a Relational Spatio-Temporal Representation

Muralikrishna Sridhar and **Anthony G Cohn** and **David C Hogg** [1]

Abstract.   We propose a framework that learns functional object-categories from spatio-temporal data sets such as those abstracted from video. The data is represented as one activity graph that encodes qualitative spatio-temporal patterns of interaction between objects. Event classes are induced by statistical generalization, the instances of which encode similar patterns of spatio-temporal relationships between objects. Equivalence classes of objects are discovered on the basis of their similar role in multiple event instantiations. Objects are represented in a multidimensional space that captures their role in all the events. Unsupervised learning in this space results in functional object-categories. Experiments in the domain of food preparation suggest that our techniques represent a significant step in unsupervised learning of functional object categories from spatio-temporal patterns of object interaction.

1   Introduction

Children learn about the world around them by observing and participating in activities that engage them in the course of every day life. One aspect of learning activity models involves acquiring notions of what objects mean to them based on the function they fulfill in activities. Functional categories and taxonomies of objects are naturally acquired by humans during the process of observing object behaviour and using them accordingly. An important step toward unsupervised learning of activity models is to learn an analogous model of functional object categories purely by observing their behaviour.

In this work, we represent the behaviour of objects involved in an activity, in terms of an *activity graph*, which captures qualitative spatio-temporal patterns of interaction between these objects. We search for frequent similar subgraph instances and generalize these by variablizing. These are our event classes, the instances of each event class encoding a similar pattern of spatio-temporal relationships between their respective object instances.

Then we learn object categories by clustering in an object space, where the similarity measure between objects is measured, based on whether they play a similar role across the event instances for each of the event classes; e.g., a set of objects, even though different in appearance, may tend to play a similar role in events such as washing, cutting and cooking as opposed to others that do not play such a role in these events. By observing multiple instances of such event classes that have the same *e*vent role for this set of objects, it is natural to form a category that correspond to what we refer to as vegetables.

Through our experiments we demonstrate that using our framework it is possible to learn semantically meaningful *functional object categories and a taxonomy* purely by observing object behaviour.

In section 3 we show how functional object categories can be learned from event classes. The rest of the paper describes a novel procedure for inducing event classes from video input.

2   Related Work

Much previous work has focused on supervised learning of object classes either based on the appearance of the object itself [9] or by recognizing contextual cues such as activities associated with objects [8] to locate and recognize objects. By contrast, unsupervised learning of objects can be divided into two stages, the first being object discovery e.g. discovery of blobs that are candidates for objects from video. The second stage is object class learning which involves automatically categorizing these blobs into object classes. Early work on object discovery [6] formed candidate objects by grouping pixels with similar temporal signatures that are constructed by recording colour (RGB) values for stable intervals when objects arrive, stay and depart from a region. In [7], candidate objects are obtained by first over segmenting images in a video and after extracting image features for these segments, rigidly moving features are grouped into potential objects.

Both object discovery and class learning is performed simultaneously from a collection of static images [5] in two steps. First multiple segmentations for each image are produced, by varying the parameters of the normalized cut technique with the assumption that each object instance is correctly segmented at least by one segmentation. Then object classes which are groups of correctly segmented objects that are coherent in a large set of candidate segments, are learned. Another approach [1] obtains a hierarchy of object classes for static scenes by grouping image features which spatially co-occur across images for the same scene, under the same leaf of the hierarchy. In this manner, the technique learns to identify candidate objects such as keyboards, while also learning higher level object classes such as a desk area (consisting of a computer, desk etc).

In this work we perform object discovery by first over segmenting the video in terms of colour patches and then grouping spatially cohesive and continuous coloured blobs to discover a candidate set of objects. We perform object class learning by clustering on a object space, where the similarity between objects is based on similar spatio-temporal behaviour (specifically object interactions) in scenes.

Recent work on event learning [3, 4] aims at learning activity/event classes given a sequence of primitive events, where the primitive events are defined and recognized a priori. In [2] a relational representation language is introduced for defining temporal events, and algorithms for learning these definitions from video output are described. In this work, we introduce a generic definition for events, in terms of graphs, that captures changing spatio-temporal
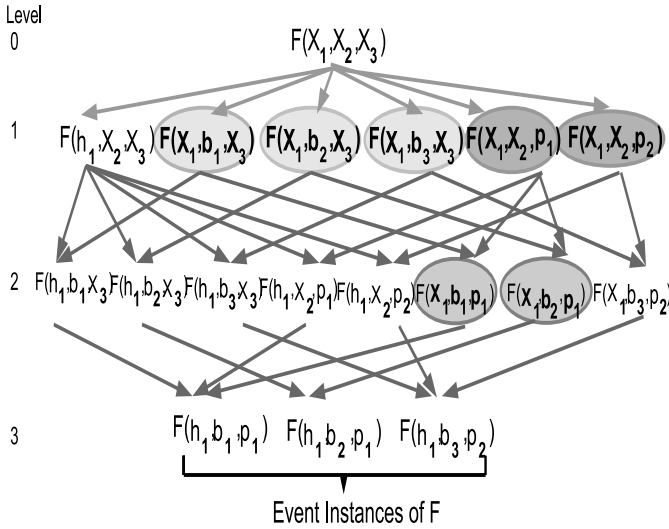
Figure 1. Lattice for general to specific object learning

relationships between discovered objects. We show how this representation enables event mining and object learning.

3 Object Learning

Assume the existence of a set of *event classes* $F(\bar{X})$, where $\bar{X}$ is a sequence of object variables in some canonical ordering, between which some set of spatio-temporal relationships hold and which when instantiated, yields a set of event instances. The event classes $F_i(\bar{X}) = F_i(X_1, .., X_k, ..., X_m)$ in general have multiple event instances in the corpus so that all these instances encode the same set (or more generally a similar set) of spatio-temporal relationships between their objects. This induces a natural mapping between objects corresponding to each object variable X_k for the event instances of an event class. Given a corpus of such instances, we show, using an example, how to induce functional object categories for the set of objects present in these instances. The event classes could be hand-crafted manually through knowledge engineering techniques, or, as we describe in later sections, could be induced from a video by an event learning procedure.

Let $F(X_1, X_2, X_3)$ be an example event class that represents events such as "X_2 being lifted away from of X_3 by X_1". The example in fig 2(c), is one such *event instance* ($F(h_1, b_1, p_1)$) of the event class F with object instances h_1, b_1, p_1 having IDs $3, 4$ and 6 respectively. Let us suppose that two other instances $F(h_1, b_2, p_3), F(h_1, b_3, p_2)$ of the same class F had been observed in the scene.

A lattice as shown in fig. 1 is grown from event instances at the bottom level (3), by generalizing exactly one argument position to a variable at each successive level. We then search for equivalence classes of objects from general to specific by traversing down this lattice, using the following procedure. For every node of each level l in the lattice, the procedure involves searching for sets of nodes at level $l + 1$, where each set is formed by substituting more than one object instance for the same variable X_k, for that node at level l.

Applying this procedure at level 0 of the lattice, we get two such sets at level 1 (shaded with two colours) : $\{F(X_1, b_1, X_3), F(X_1, b_2, X_3), F(X_1, b_3, X_3)\}$ obtained by substituting for X_2 with b_1, b_2, b_3 and

$\{F(X_1, X_2, p_1), F(X_1, X_2, p_2)\}$ obtained by substituting for X_3 with p_1, p_2 respectively . As the substituted constants $\{b_1, b_2, b_3\}$ and $\{p_1, p_2\}$, play the same roles (as the variables X_2 and X_3 respectively) for the event class F, we say that F has induced *event roles* for instances of the variables X_2 and X_3 resulting in equivalence classes $\{b_1, b_2, b_3\}$ and $\{p_1, p_2\}$ respectively.

We now show that, by applying the same procedure at one level below (level 1) of the lattice, we obtain a more specific event role for the specific event of *objects placed on a certain plate (p1)*. The procedure applied at level 1 results in a set of nodes $\{F(X_1, b_1, p_1), F(X_1, b_2, p_1)\}$ at level 2 (as shaded in fig. 1), obtained by substituting for X_2 in $F(X_1, X_2, p_1)$ with b_1, b_2 respectively. We say that the more specific event class $F(X_1, X_2, p_1)$ has induced a more *specific event role* for the variable X_2 resulting in an equivalence class of objects $\{b1, b2\}$, i.e. objects being put on plate p_1. By progressively traversing down the lattice using this procedure, it becomes possible to create event roles and corresponding equivalence classes $C_1...C_n$, from general to specific.

Applying this idea, we produce a matrix of *object by equivalence classes*, O in which $O_{i,j}$ equals 1 if the object i occurs in the equivalence class C_j and 0 otherwise. As each equivalence class corresponds to an event role, the row vectors of this matrix summarize each object in terms of the role it plays in all the event-roles and thus induce a multidimensional *object space*. In this space, objects that have a similar role with respect to similar sets of events are expected have a high similarity measure. We therefore perform k-means clustering using a cluster partition index to determine k. Hierarchical clustering on these categories then yields an object taxonomy.

In the next section, we show how event classes can be learned from video input and in section 6 the results of applying our object learning procedure are discussed.

4 Activity Graphs from Video

Object discovery is performed by first over segmenting the video in terms of colour patches and then grouping these into *spatially continuous and cohesive* blobs that are a mix of noisy patches along with potential objects. These blobs are given IDs and their position and extent are recorded from the video.

The spatio-temporal patterns in the entire video are represented using an *activity graph*. The spatial relationships between the bounding boxes of each pair of objects for every frame are mapped to a set of spatial primitives $\Re = \{\mathsf{D}, \mathsf{S}, \mathsf{T}\}$. Two objects are either spatially Disconnected(D) or connected through the Surrounds(S) or Touches(T) relationships [2]. illustrated in fig. 2(b).

For each pair of objects, these spatial relationships hold during a time interval. In general, If $\{o_1, o_2...o_n\}$ is the set of all the objects observed in the video, for each pair o_i, o_j, a particular spatial relationship $r \in \Re$ holds for each frame f, i.e. $holds(r(o_i, o_j), f)$. We are interested in maximal one-piece time intervals during which r holds between o_i and o_j, which we refer to as *episodes*.

We represent such episodes by a quadruple $E = \langle o_i, o_j, \tau, r \rangle$, where $|\{r : \text{Holds}(r(o_i, o_j), f) \in \tau\}| = 1$ and τ is a consecutive sequence of frames such that $\forall \tau' (\tau \subset \tau' \rightarrow |\{r : \text{Holds}(r(o_i, o_j), f) \in \tau'\}| > 1$. We thus obtain the set of all episodes $\Delta = \{E_1, E_2...E_m\}$ for all pairs of objects. Episodes labelled $E_1 - E_{20}$ in fig 2(a) correspond to this set, for the activity considered in this example.
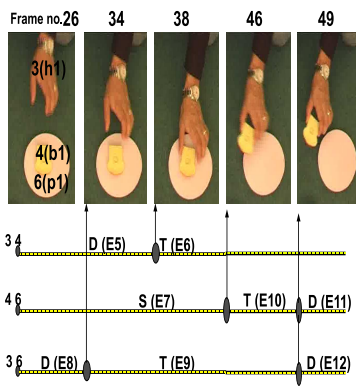
---

[2] This approach clearly could be applicable to any set of spatial relations $\Re'$. Our simplified approach to video analysis is 2D, thus using this set of spatial relations means, e.g. an object o_1 placed on an object o_2 is represented as $\mathsf{S}(o_1, o_2)$ – these 3 relations have sufficed for our experiments.
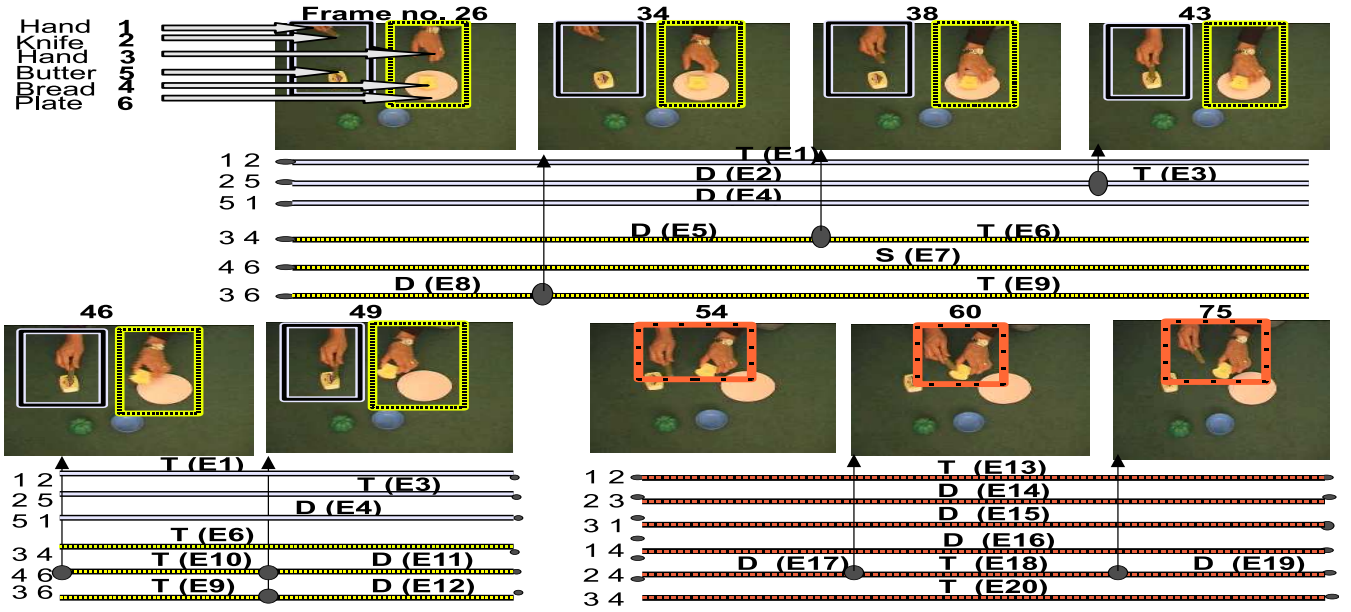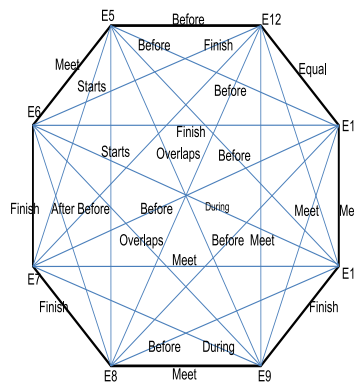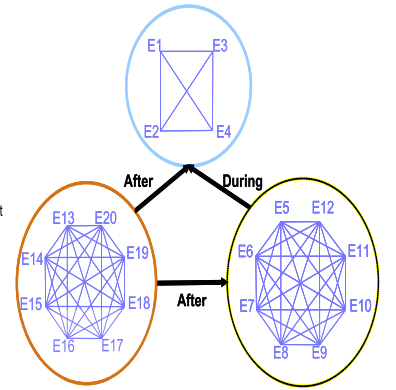
(a) An activity
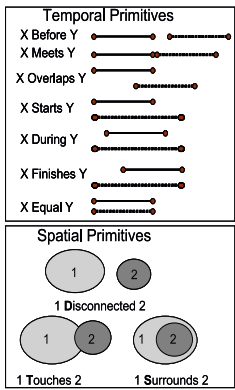
(b) Spatial and Temporal Primitives

(c) A subactivity of the activity in (a)

(d) Level-0 activity graph for episodes $E_5 - E_{12}$ in (c)

(e) Level-1 Activity Graph for episodes $E_1 - E_{20}$ in (a)

Figure 2.

Having obtained all the episodes, we obtain a complete graph – which we call an *activity graph* – whose vertices represent the episodes and whose edges relate the time intervals corresponding to their respective episodes using Allen's temporal primitives $\Im$. We call the complete graph encoding all temporal relationships between intervals $E_1 - E_{20}$ a *level-0 activity graph* for the activity in fig. 2(a).

More formally, we have the activity graph $(V, E, \eta, \rho, \Delta, \Im)$, where the function $\eta : V \longrightarrow \Delta$ maps the vertices $V = \{v_i\}$ to episodes in Δ and $\rho : E \longrightarrow \Im$ maps the directed edges between all pairs of vertices $E : e_{ij} = \langle v_i, v_j \rangle$ to temporal relationships in $\Im$. We require that η is a bijective mapping from vertices to the set of episodes in the activity graph.

The complete activity graph is too large to display here and a typical activity graph is too complex to be able to search to find event

classes[3]. Fig. 2(d) shows a subgraph of the level-0 activity graph for episodes $E_5 - E_{12}$ - depicted in fig. 2(c). Therefore, prior to searching for event classes we use an attention mechanism to structure and simplify the level-0 activity graph to produce a level-1 activity graph. This is achieved by using a foreground attention mechanism (described below) to cluster episodes and forming a new graph structure over these clusters. Each cluster represents an atomic event and we call the clusters of episodes and their Allen relationships, a *unary event graph* (unary EG). The graph whose nodes are unary event graphs and whose edges are Allen's temporal relationships between these nodes is the *level-1 activity graph*.

---

[3] If we consider $n = 10$ objects and k as the average number of episodes in video which is usually 10^2 even for scenes that last for a minute, the activity graph results in a search space of $O(k^2 n^4)$ .i.e $O(10^8)$.

Foreground Attention Mechanism: We hypothesize that many activities can be conceived in terms of different foreground events each of which involve interactions only between a subset of objects – *foreground objects*, at different time periods. This idea can be intuitively explained using fig. 2(a), where the entire activity shown can be conceived in terms of three *foreground events* - (1) the *left hand* scooping some *butter* with a *knife* (2) the *right hand* removing the *bread* from the *plate* (3) the *left hand* spreading *butter* on the *bread* with a *knife*, while the *right hand* holds the *bread*.

As long as {*left hand*,*knife*, *butter*} and {*right hand*, *plate* , *bread* }, are disconnected, we have two sets of foreground objects $\{1, 2, 5\}$, $\{3, 4, 6\}$, between frames 26 and 49. When the knife and the bread start to interact, the foreground set changes to the set of IDs $\{1, 2, 3, 4\}$, in which the butter and plate with IDs 5 and 6 are not included (frames 54-75). Three periods and their corresponding set of episodes $\{E_1 - E_4\}$, $\{E_5 - E_{12}\}$, $\{E_{13} - E_{20}\}$ (as shown in the parallel lines below the frames), for the three foreground events are thus obtained. The next two paragraphs describe how, in general foreground events are detected and may be ommited on a first reading.

We look for spatial changes between a pair of objects. For each such pair of *primary foreground objects* o_1, o_2 at some frame f, we find the set Ω of all *moving* objects which are connected (i.e. T or S) to o_1 or o_2, or which are connected to o_1 or o_2 indirectly via another *moving* object which is connected to o_1 or o_2 (directly or indirectly). The set Ω is propagated forwards to some frame f_2 and backwards to some frame f_1 from f until such time that one of the objects in $\Omega - \{o1, o2\}$ (the *secondary foreground objects*) changes its spatial relation to some other object in Ω to D, (unless o_1 and o_2 are connected at that time). The entire time from f_1 to f_2 is termed a *period* during which a *foreground event* involving o_1 and o_2 occurs, involving all the foreground objects Ω.

The intuition behind this definition is that a spatial change focuses attention on a pair of objects (at least one of which must be moving, since a change has occurred), and all the objects which are intimately connected to the two objects, and groups all the interactions involving the primary objects together until such time as one of the secondary objects becomes fully disconnected from the group of objects (which then terminates this particular set of foreground objects). Note that it is possible, depending on the choice of primary objects o_1 and o_2 for there to be multiple temporally overlapping foreground events involving shared objects (though this has not occurred in the videos we have analysed so far).

For each foreground event, we create a unary event graph (unary EG) restricted to the foreground objects of the foreground event and just during the temporal extent of the foreground event. Each unary EG endures for a period P and can be represented by the unary EG $(V, E, \eta, \rho, \Delta_P, \Im)$ between the episodes for the time period P. The three unary EGs for the activity in fig. 2(a) are shown as the nodes in the level-1 activity graph in fig. 2(e). Unary EGs (which are single nodes of the level-1 activity graph) typically capture simple events such as removing a slice of bread from a plate.

In the next section we show how to generalize unary events to unary event *c*lasses, and then how to form n-ary event classes, which are compound event classes composed of unary event classes. Instances of n-ary event classes are n-ary events which are composed of n unary EGs of the level-1 activity graph and which represent complex events such as the entire activity depicted in fig. 2(a,c).

5 Event Learning

The activity graph consists of many individual events; these can be similar in that they have similar spatio-temporal relationships between their constituent objects. In order to formalize the idea of an *event class* that captures these regularities, independent of the actual objects involved, we first introduce a generalized version of an *unary event graph*. We then show how n-ary event classes can be formed, consisting of individual unary event classes.

To generalize events to *event classes*, we first consider a unary EG $\phi = (V, E, \eta, \rho, \Delta_P, \Im)$ for a time period P. Instead of object instances $o_i \in \Omega$ and intervals $\tau \in \Lambda$, consider sets of object and interval *variables* $X = \langle X_O, X_T \rangle$ so that $O_i \in X_O$ and $T \in X_T$[4]. We can now generalize the set of episodes $E \in \Delta_P$ to $E_X \in \Delta_X$ where Δ_X is a set such that $E_X \in \Delta_X$ if and only if $E_X = \langle O_1, O_2, T, r \rangle$ where $O_1 \in X_O, O_2 \in X_O, T \in X_T, r \in \Re$. We use the generalised set of episodes to formalise event classes by first defining a *unary event class graph* (unary ECG) which captures a common pattern of spatio-temporal relationships amongst a set of similar unary EG (instances), in a generic form.

Definition Let $\phi = (V, E, \eta, \rho, \Delta_P, \Im)$ be a unary EG of the transformed activity graph, then $\gamma = (V', E', \eta', \rho', \Delta_X, \Im)$ is a *unary event class graph (unary ECG)* of ϕ, or we say that γ *θ-generalizes* ϕ if $\exists \theta = \theta_O \cdot \theta_T$ where $\theta_O : X_O \rightarrow \Omega$ and $\theta_T : X_T \rightarrow \Lambda$, such that γ is isomorphic to ϕ under the substitution θ, i.e.

1. $\{\eta'(v')\theta : v' \in V'\} = \{\eta(v) : v \in V\}$.
2. $\{\rho'(e'_{ij}) : e'_{ij} = (v'_i, v'_j) \in E'\} = \{\rho(e_{ij}) : e_{ij} = (v'_i\theta, v'_j\theta) \in E\}$.

We require that a unary ECG generalises at least λ unary EGs, i.e. instances must occur frequently.

We now extend the the idea of a unary event class graph to an n-ary event class graph (n-ary ECG) composed of unary ECGs. A n-ary ECG is just a graph made up of unary ECGs $\gamma_1...\gamma_n, n > 2$ as its vertices and whose edges relate the time periods P_i and P_j corresponding to γ_i and γ_j by Allen's temporal primitives $\Im$. A n-ary ECG Γ whose vertices are the set $\{\gamma_1, ..., \gamma_n\}$ θ-generalizes an n-ary EG Φ with vertices $\{\phi_1, ..., \phi_m\}$, if each γ_i θ-generalizes a corresponding ϕ_i and the temporal relationship between any $\langle \phi_i, \phi_j \rangle \in \Phi$ is the same as for the corresponding $\langle \gamma_i, \gamma_j \rangle \in \gamma$. A n-ary ECG represents a *n-ary event class* if it generalises at least λ n-ary EGs. We model λ as an exponential decreasing function of n in order to allow for larger n-ary ECGs to θ-generalise fewer n-ary EGs.

Using these definitions, we finally formalize *event classes* as maximal event class graphs. We define a maximal event class graph (MECG) as a event class graph which generalises some set of EGs, such that no other ECG which contains it generalizes this set. I.e. every MECG generalises a set of EGs which are not generalised by some larger ECG. The procedure for computing MECGs involves two stages. In the first stage, unary ECGs with a statistically significant number of EG instantiations are found. In the second stage, these unary ECGs are iteratively used to build larger and larger ECGs (with statistically significant number of instantiations), until a final set of MECGs are obtained. In this manner we discover event classes as MECGs from the level-1 activity graph.

Having found all the MECGs, we give them names $F_1(\bar{X})...F_k(\bar{X})$, where $\bar{X}$ is a sequence of variables in the

---

[4] Note that we use capitalized/bold letters for variables and small letters for instances.
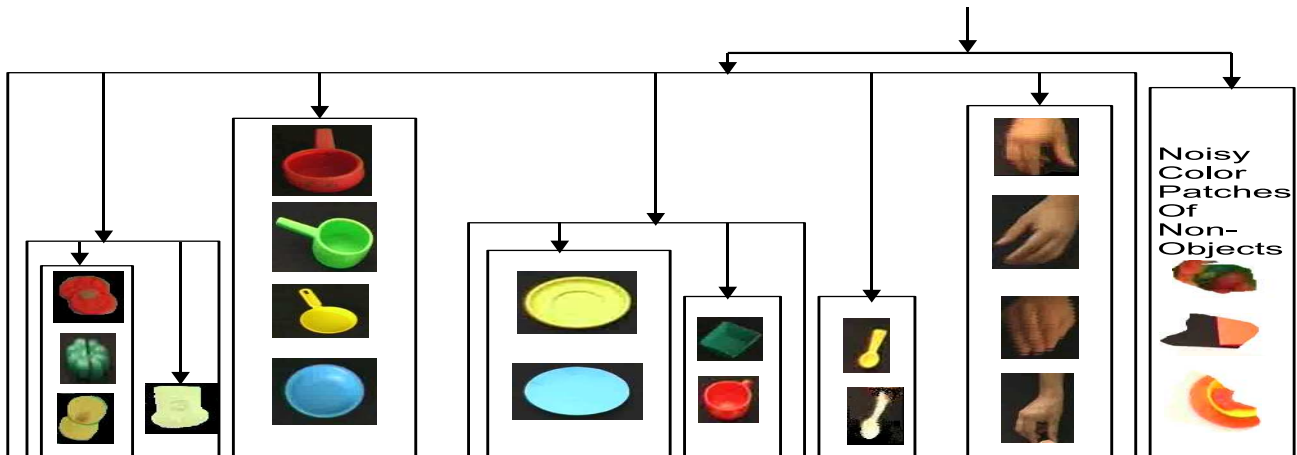
Figure 3. A hierarchy of objects categories.

MECGs, in some canonical ordering of nodes in each MECG. In section 3, where we were purely concerned with inducing an object taxonomy from the event definitions we ignored the internal structure of an MECG and used just these $F_i(\bar{X})$, which can be defined as predicates from each of the MECGs.

6 Experiments

We demonstrate our framework using a video taken with a toy (plastic) kitchen set up. We have chosen a constrained environment for the moment, in order to minimize the complexities arising in a real kitchen as a result of cluttered backgrounds, flickering lights, shiny surfaces, multiple shadows etc. We have further simplified the problem by focusing only on the hand (not the entire person) along with the other objects in the kitchen scene and taking care in the actions of the cook to not create complications arising, for instance, from full occlusion of objects involved. However, despite such simplifications, a large number of noisy patches are produced from the object discovery module, making the learning problem challenging. The video is taken with a static overhead camera that focuses on the scene. The scene consists of hands simulating the preparation of sandwiches, hot drinks, cutting vegetables and cooking vegetable dishes, lasting around 10 minutes. The video consists of exactly one instance for each of these preparations.

After applying event and object learning, we obtain the object hierarchy in fig. 3. While our procedure outputs a hierarchy of object IDs, we replace these labels with the corresponding objects from the video, in order to visualize the results. It can be observed that the proposed framework has been able to differentiate between broader categories such as food items and containers and interestingly separate noisy patches from all other objects. Finer levels of granularity are captured in the grouping which separates a slice of white bread from another group consisting of vegetables. A distinction between plates pans and spoons is also clear from the hierarchy. It can therefore be concluded that the learned categories and taxonomy is intuitive and corresponds to a functional classification of objects.

7 Summary and Future Work

A framework for learning object and event categories from video has been introduced. This framework offers a general way of representing activities in terms of spatio-temporal graphs. Techniques for mining events from this graph and then learning object functional categories from these events have been proposed in this work. Our experiments show that our framework offers a promising approach toward learning functional categories.

In the future, we plan to extend this framework in several directions. At present, event generalisation requires exact graph isomorphism. We plan to extend event classes to generalize a larger set of event instances by experimenting with similarity metrics between our event graphs. This will allow our approach to exploit a greater variety of video input to learn event and object taxonomies , and to cope better with noise (which might also intervene during an event instance). In contrast to almost all work in object recognition which is based on learning categories based on perceptual features, we have tackled the little researched problem of learning categories from function. However, there is clearly scope to use the learned functional categories to supervise visual appearance based object learning.

REFERENCES

[1] D.Parikh and C. Tsuhan;, 'Unsupervised learning of hierarchical semantics of objects (hsos).', *Computer Vision and Pattern Recognition, 2007. CVPR '07.*, 1–8, (2007).

[2] Givan R.L. Fern, A.P. and J.M. Siskind, 'Specific-to-general learning for temporal events with application to learning event definitions from video', *Artificial Intelligence Research (JAIR)*, **17**, 379–449, (2002).

[3] Somboon Hongeng, 'Unsupervised learning of multi-object event classes', *in: Proc. 15th British Machine Vision Conference (BMVC'04), London, UK, 2004*, 487–496, (2004).

[4] A. Bobick R. Hamid, S. Maddi and I. Essa, 'Structure from statistics - unsupervised activity analysis using suffix trees', *in Proc. of Conf. on Computer Vision*, (2007).

[5] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman, 'Using multiple segmentations to discover objects and their extent in image collections', *CVPR 06: Proc. of Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, 1605–1614, (2006).

[6] Brandon C.S. Sanders, Randal C. Nelson, and Rahul Sukthankar, 'A theory of the quasi-static world', *Proc. 16th Int'l Conf. on Pattern Recognition ICPR02*, (2002).

[7] Tristram Southey and James J. Little, 'Object discovery using motion, appearance and shape', *Cognitive Robotics Workshop, AAAI*, (2006).

[8] M Veloso, P Rybski, and F von Hundelshausen, 'Focus: A generalized method for object discovery for robots that observe and interact with humans', *Proc. Conf. on Human-Robot Interaction*, (2006).

[9] A. Rosenfeld P.J. Phillips W. Zhao, R. Chellappa, 'Face recognition: A literature survey', *ACM Computing Surveys*, 399–458, (2003).